

**Initial Report on Phase Two  
of the uniTEST Validity Study**

**October 2006**



# Contents

|   | <b>page</b> |
|---|-------------|
| <b>Summary/Comments</b>                               | <b>3</b>    |
| <b>Test structure and construct</b>                   | <b>4</b>    |
| <b>Demographic characteristics of the 2006 sample</b> | <b>5</b>    |
| <b>Test score scales and distributions</b>            | <b>9</b>    |
| <b>Analysis of the background of the sample</b>       | <b>14</b>   |
| <b>Person-item analysis</b>                           | <b>21</b>   |
| <b>Differential gender performance</b>                | <b>31</b>   |
| <b>Factor analysis</b>                                | <b>32</b>   |
| <b>Measurement error</b>                              | <b>37</b>   |

## Summary/Comments

- The 2006 cohort was much more able in general than the 2005 cohort, with a difference in mean score of around one standard deviation on each scale (this difference in ability was reflected in the mean GCSE scores of the two groups). In 2005, students were recruited from a representative sample of UK schools and colleges. In 2006 students were recruited from a pool of applicants to courses nominated by participating universities, many of which are highly selective.
- The 2006 cohort student scores were well spread for each component and for the total test, though more very difficult items may be useful in future for groups as strong as the 2006 group.
- The basic test statistics, such as reliability, were good (internal consistency for each component was around 0.8, and total test over 0.9).
- The inter-component correlation pattern was much as expected in relation to the construct.
- Factor analysis confirms that a three-factor solution matching the current three components is satisfactory, though the test could be used in a two factor format or might be divided in other ways.
- Empirical research is needed to identify the best way of optimising the test design and component use to maximise prediction of course success.
- Analysis indicates that differential gender performance is not due to any significant test bias, but reinforces that test components should be used differentially for different course selections.
- Measurement error and item difficulty distributions need to be considered in light of future test uses, the ability ranges of candidates who will use it and likely selection cut-off scores. It may be that different linked versions of the test will be useful if the test is required for a broad range of student abilities and a variety of cut-off scores.
- Differences in test scores between attainment groups are statistically significant. Students with high prior attainment tend to obtain high uniTEST scores. There is, however, some evidence that the test reveals academic potential in some candidates that, for whatever reason, have not done well in their GCSEs. It is reasonable then to assume that for some candidates their reasoning abilities are not reflected in their GCSE results.
- Differences in test scores between field of study groups are statistically significant. The students who obtained the highest scores in the test are those applying for degrees in the fields of Engineering/Architecture and Computers/IT. The lowest scores are obtained by students wanting to pursue degrees in the field of Education/Social and Nursing. (It must be noted that these differences may be affected by the dominance of some subjects by applicants to certain universities.)

- Students from socially deprived areas have, on average, lower scores than do their counterparts from more advantaged areas. However, there are some students from deprived areas who get scores that are over the average mark in the uniTEST.
- The background variable that has the largest positive relationship with uniTEST scores is candidates' prior attainment (mean GCSE).
- In comparison to the 'White' group, other ethnic groups such as Bangladeshi, Black African, Chinese and Indian, generally did not perform as well in the test.
- After allowing for candidates' performance at GCSE, deprivation, number of people in the area that have level 4/5 qualifications, distance travelled to work and number of lone parent households with dependent children do not seem to be associated with students' success in the uniTEST. The effects of centre type, students' disabilities and socio-economic group were also not significant.

## Test Structure and Construct

uniTEST focuses on particular skills important in higher education. It assesses a student's capacity to reason in a range of familiar and less familiar contexts which do not require subject specific knowledge. It is expected that the wider the range of contexts in which a student is able to reason, the more successful they are likely to be in applying these skills in new contexts and future study.

There are three components to uniTEST: Quantitative Reasoning (dealing with information and problem solving), Critical Reasoning (decision making and argument analysis) and Verbal and Plausible Reasoning (interpretation and socio-cultural understanding). All questions are in multiple choice format.

Quantitative Reasoning is reasoning typically, but not exclusively, in the domains of mathematics and science, and includes the application of generally accessible quantitative, scientific and technological information.

Verbal and Plausible reasoning is the kind of reasoning typical in the arts, humanities and social sciences, including verbal and visual comprehension, holistic judgements about meaning, and socio-cultural understandings (e.g. the interpretation of subjective human constructs).

Critical Reasoning addresses general reasoning in both broad domains, and is relevant to a range of courses including scientific, technical, business, humanities and social sciences.

Each component had 30 core items, and all core items were given to all students. The CR component was subdivided into three parts, CR1, CR2 and CR3. Although these had a similar common focus, CR1 tended to focus on non-text based, decision making material requiring a degree of deductive reasoning, whereas CR3 was all text-based with a Humanities issues focus requiring a degree of inductive reasoning. CR2 bridged the gap between these, being text based but requiring both deductive and inductive reasoning.

Each test paper also included five trial items for statistical evaluation for future use. Performance on these items did not contribute to the students' scores.

## Demographic characteristics of the 2006 sample

The uniTEST was taken in May and June 2006 by 1589 students who had applied to one or more of the seven universities participating in the validity study. Table 1 shows the institutions taking part in the study, their ranking and the number of students applying to them who have taken the uniTEST.

**Table 1: Institutions taking part in the validity study**

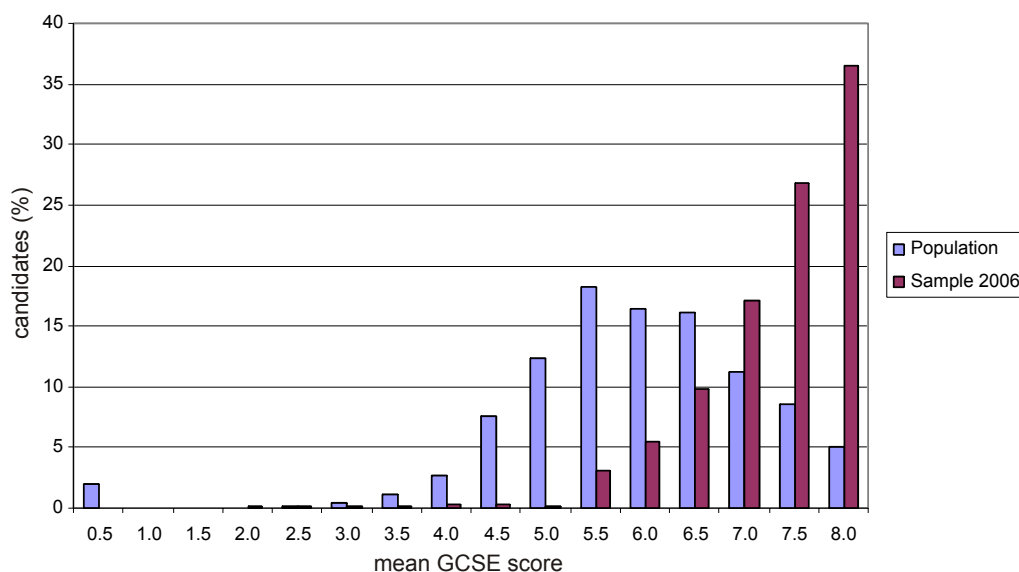
| Institution                 | Ranking | Number of students | Percentage of students |
|-----------------------------|---------|--------------------|------------------------|
| University of Bristol       | 9       | 214                | 13.5                   |
| University of Cambridge     | 2       | 702                | 44.2                   |
| University of Durham        | 10      | 205                | 12.9                   |
| University of Edinburgh     | 14      | 120                | 7.6                    |
| University of Exeter        | 34      | 188                | 11.8                   |
| University of Hertfordshire | 58      | 64                 | 4.0                    |
| University of Warwick       | 8       | 413                | 26.0                   |

It should be noted that the 'number of students' and percentage columns sum to more than 1589 and 100 respectively. This is because some students had applied to more than one participating institution.

The majority of these universities are ranked quite highly (according to the Times Top 100 Universities) and, therefore, the uniTEST sample is biased towards high attainers.

The graph in Figure 1 shows the distribution of the mean GCSE for the general population and the 2006 uniTEST sample. The mean GCSE was computed after converting the grades into scores following the UCAS tariff (A\*=8, A=7, B=6, etc). For the population, this mean was computed using the GCSE scores of students who were 17 years old and had AS/A level results in 2005 (215948 students). For the sample, we took into account the GCSE results of the students that participated in the study and have a match in the 16+/18+ databases (1170 students).

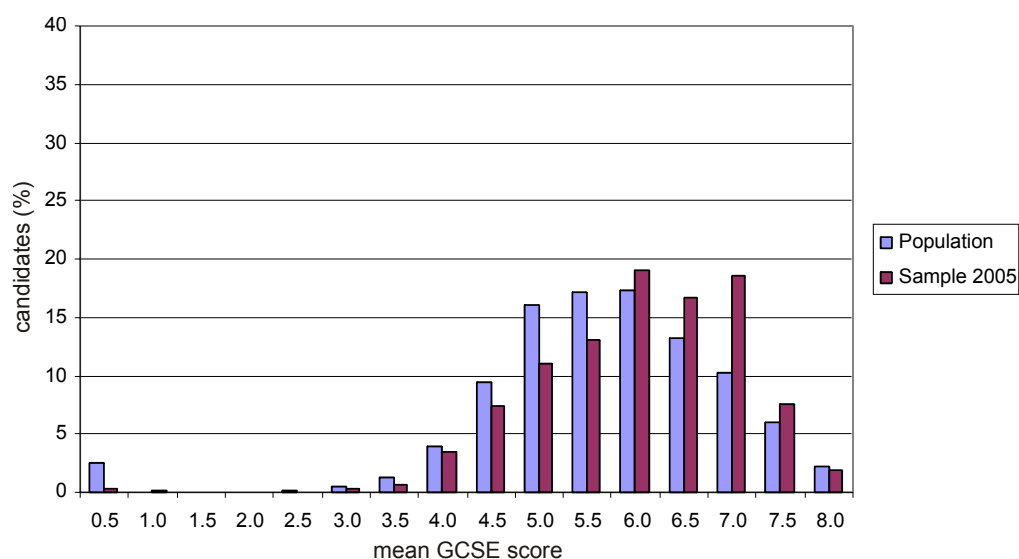
**Figure 1: Distribution of mean GCSE scores for the general population and the 2006 uniTEST sample**



Students taking the test in 2006 were recruited from the pools of applicants to courses nominated by the participating universities, most of which are highly selective. This is reflected in the very high mean GCSE scores of the sample when compared to the population of all 17 year-old AS/ A level students.

The graph in Figure 2 shows a similar comparison of the mean GCSE distribution for the general population and the 2005 uniTEST sample. In 2005, students were recruited from a representative sample of UK schools and colleges, resulting in a sample whose mean GCSE distribution more closely matches that of the population as a whole.

**Figure 2: Distribution of mean GCSE scores for the general population and the 2005 uniTEST sample**



The 2006 sitting, to date, consisted of 1589 candidates, 937 females (59%) and 652 males (41%).

**Table 2: Gender of the 2006 uniTEST Candidates**

| Gender | Frequency | Percent | Cumulative percent |
|--------|-----------|---------|--------------------|
| Female | 937       | 59.0    | 59.0               |
| Male   | 652       | 41.0    | 100.0              |
| Total  | 1589      | 100.0   |                    |

Most of the candidates (94.8%) were born in 1987 and 1988.

**Table 3: Years of Birth of 2006 uniTEST candidates**

| Year of Birth | Frequency | Percent | Cumulative percent |
|---------------|-----------|---------|--------------------|
| 1957          | 1         | 0.1     | 0.1                |
| 1958          | 1         | 0.1     | 0.1                |
| 1968          | 1         | 0.1     | 0.2                |
| 1975          | 1         | 0.1     | 0.3                |
| 1981          | 1         | 0.1     | 0.3                |
| 1982          | 2         | 0.1     | 0.4                |
| 1983          | 2         | 0.1     | 0.6                |
| 1984          | 4         | 0.3     | 0.8                |
| 1985          | 8         | 0.5     | 1.3                |
| 1986          | 38        | 2.4     | 3.7                |
| 1987          | 549       | 34.6    | 38.3               |
| 1988          | 957       | 60.2    | 98.5               |
| 1989          | 23        | 1.4     | 99.9               |
| 1990          | 1         | 0.1     | 100.0              |
| Total         | 1589      | 100.0   |                    |

Most students were from Comprehensive schools (34%) and Independent schools (32%). The rest came from a variety of institution types as indicated.

**Table 4: School Types of 2006 uniTEST Candidates**

| Institution type       | Frequency | Percent | Cumulative percent |
|------------------------|-----------|---------|--------------------|
| Art Design and Per Art | 34        | 2.1     | 2.1                |
| Comprehensive School   | 541       | 34.0    | 36.2               |
| Further Education      | 14        | 0.9     | 37.1               |
| Grammar School         | 135       | 8.5     | 45.6               |
| Grant Main Sec (State) | 171       | 10.8    | 56.3               |
| Independent School     | 510       | 32.1    | 88.4               |
| Other Secondary School | 3         | 0.2     | 88.6               |
| Sixth Form Centre      | 17        | 1.1     | 89.7               |
| Sixth Form College     | 163       | 10.3    | 99.9               |
| Tertiary College       | 1         | 0.1     | 100.0              |
| Total                  | 1589      | 100.0   |                    |

For the purposes of analysis, students were assigned to Field of Study Groups, based on the degree courses to which they had applied.

**Table 5: Field of Study Groups of 2006 uniTEST Candidates**

| Field of Study           | Frequency | Percent | Cumulative percent |
|--------------------------|-----------|---------|--------------------|
| Arts/Humanities          | 496       | 31.2    | 31.2               |
| Business/Commerce        | 202       | 12.7    | 43.9               |
| Computers/IT             | 16        | 1.0     | 44.9               |
| Engineering/Architecture | 79        | 5.0     | 49.9               |
| Education/Social         | 11        | 0.7     | 50.6               |
| Law/Legal                | 135       | 8.5     | 59.1               |
| Medicine/Dentistry       | 117       | 7.4     | 66.5               |
| Nursing                  | 14        | 0.9     | 67.3               |
| Science/Maths            | 519       | 32.7    | 100.0              |
| Total                    | 1589      | 100.0   |                    |

With respect to ethnicity, about 78% reported to be white. Most of the rest reported being Asian. Less than 2% reported being Black.

**Table 6: Ethnicity of the 2006 uniTEST Candidates**

| Ethnicity             | Frequency | Percent | Cumulative percent |
|-----------------------|-----------|---------|--------------------|
| Asian – Bangladeshi   | 7         | 0.4     | 0.4                |
| Asian – Chinese       | 41        | 2.6     | 3.0                |
| Asian – Indian        | 52        | 3.3     | 6.3                |
| Asian – Other         | 20        | 1.3     | 7.6                |
| Asian – Pakistani     | 14        | 0.9     | 8.4                |
| Black – African       | 24        | 1.5     | 9.9                |
| Black – Caribbean     | 4         | 0.3     | 10.2               |
| Black – Other         | 1         | 0.1     | 10.3               |
| Not given             | 132       | 8.3     | 18.6               |
| Not given (Dom=Osea)  | 4         | 0.3     | 18.8               |
| Other                 | 9         | 0.6     | 19.4               |
| Other Mixed           | 6         | 0.4     | 19.8               |
| White                 | 1246      | 78.4    | 98.2               |
| White and Asian       | 21        | 1.3     | 99.5               |
| White/Black African   | 3         | 0.2     | 99.7               |
| White/Black Caribbean | 5         | 0.3     | 100.0              |
| Total                 | 1589      | 100.0   |                    |

## Test score scales and distributions

Scaled scores are reported for each candidate on each of the three components and on the total test. Candidates who responded correctly to all the items in a scale are given a scaled score of 100.

Based on the raw scores, the analysis software, QUEST, gives an estimate of the candidate's ability in *logits* for each scale and for the total test. To produce the scaled scores, the logit values of the 2006 test were adjusted to the 2005 scale. (The scale scores were calculated from the adjusted logits using the same formulae as in 2005. The formulae in 2005 were derived to standardise the scaled scores to a mean of 50 and a standard deviation of 12). These calculation procedures enable direct comparison of the 2005 and 2006 scores.

The following tables give comparisons of the scaled scores in 2005 and 2006 showing that the 2006 group was much stronger on average than the 2005 group.

**Table 7: Scale Score Statistics, uniTEST 2005**

|                | Total Scale Score | Verbal-Plausible Scale Score | Quantitative Reasoning Scale Score | Critical Reasoning Scale Score |
|----------------|-------------------|------------------------------|------------------------------------|--------------------------------|
| N              | 852               | 852                          | 852                                | 852                            |
| Mean           | 50.00             | 50.00                        | 50.00                              | 50.00                          |
| Std. Deviation | 12.00             | 12.00                        | 12.00                              | 11.98                          |
| Range          | 80.88             | 78.90                        | 75.84                              | 93.34                          |
| Minimum        | 17.13             | 20.83                        | 18.98                              | 6.66                           |
| Maximum        | 98.01             | 99.73                        | 94.82                              | 100.00                         |

**Table 8: Scale Score Statistics, uniTEST 2006**

|                | Total Scale Score | Verbal-Plausible Scale Score | Quantitative Reasoning Scale Score | Critical Reasoning Scale Score |
|----------------|-------------------|------------------------------|------------------------------------|--------------------------------|
| N              | 1589              | 1589                         | 1589                               | 1589                           |
| Mean           | 60.69             | 58.83                        | 58.61                              | 62.24                          |
| Std. Deviation | 12.21             | 12.66                        | 13.57                              | 12.19                          |
| Range          | 77.56             | 78.65                        | 79.17                              | 76.10                          |
| Minimum        | 22.44             | 21.35                        | 20.83                              | 23.90                          |
| Maximum        | 100.00            | 100.00                       | 100.00                             | 100.00                         |

The difference between the mean scaled scores for the total test is 10.69 scale points, about one standard deviation. Of the individual components, the largest difference is in CR where the difference in means is equal to 12.24 scale points and the smallest difference is for QR where the difference between means is 8.61 scale points.

Of the 1589 students, 6 answered all 30 VP items correctly; 22 all QR items, and 6 all CR items.

## Correlations between scales

The product moment correlations of the scaled scores are given in the following table.

**Table 9: Correlations**

|    |                     | VP     | QR     | CR     |
|----|---------------------|--------|--------|--------|
| VP | Pearson Correlation | 1      | .499** | .654** |
|    | Sig. (2-tailed)     |        | .000   | .000   |
|    | N                   | 1589   | 1589   | 1589   |
| QR | Pearson Correlation | .499** | 1      | .662** |
|    | Sig. (2-tailed)     | .000   |        | .000   |
|    | N                   | 1589   | 1589   | 1589   |
| CR | Pearson Correlation | .654** | .662** | 1      |
|    | Sig. (2-tailed)     | .000   | .000   |        |
|    | N                   | 1589   | 1589   | 1589   |

\*\* Correlation is significant at the 0.01 level (2-tailed).

The highest correlation is that between QR and CR scaled scores with a correlation of 0.662. This means that about 44% of the variation in the scores on one scale can be explained by the scale scores in the other, the rest being due to unique features of the scale dimension plus error of measurement.

The correlation between VP and the CR scaled scores is of similar magnitude at 0.654 (about 43% common variance).

The lowest correlation of 0.499 is between the VP and the QR (about 25% common variance).

Thus, the CR component, while having distinctive features, has significant commonality with the other two components, while the VP and QR components have only a small degree of commonality.

This correlation pattern is as expected, with the CR component focusing on core reasoning skills common to many courses and bridging the gap between quantitative courses and arts/humanities courses.

More is said about the validity of the component structure in the section on factor analysis.

## Scale Score Distributions

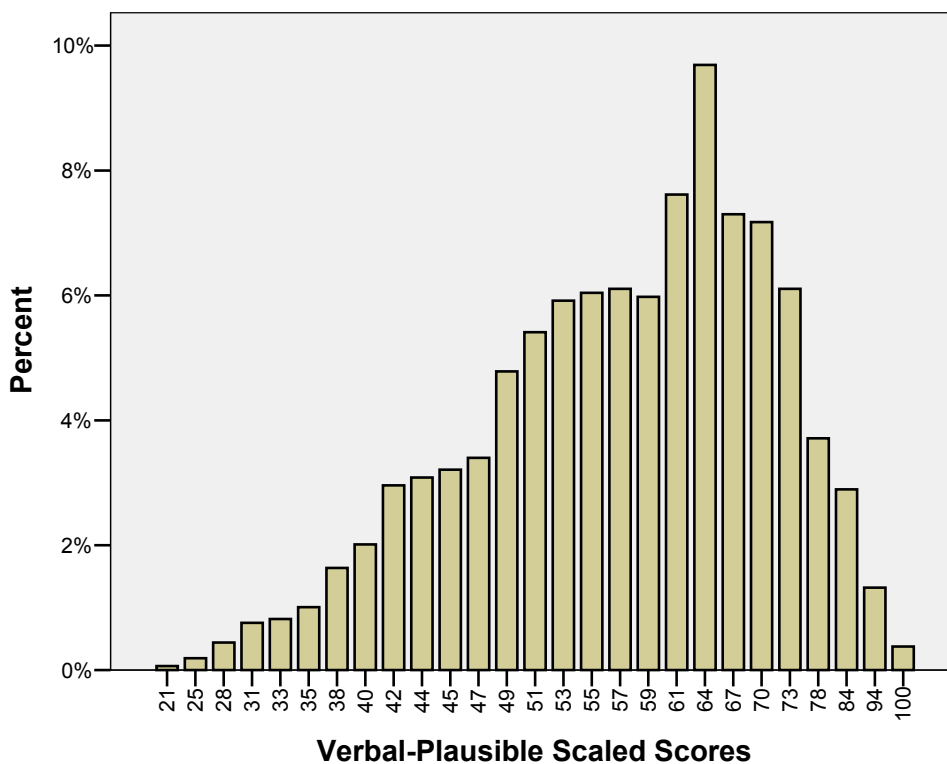
The diagrams that follow show the distributions of the scaled scores for 2006. The horizontal axis indicates increasing scaled score values from left to right; the vertical axis indicates the percentage of candidates achieving each score.

The shapes of the bar charts show that in general candidates in 2006 obtained higher scaled scores than those in 2005.

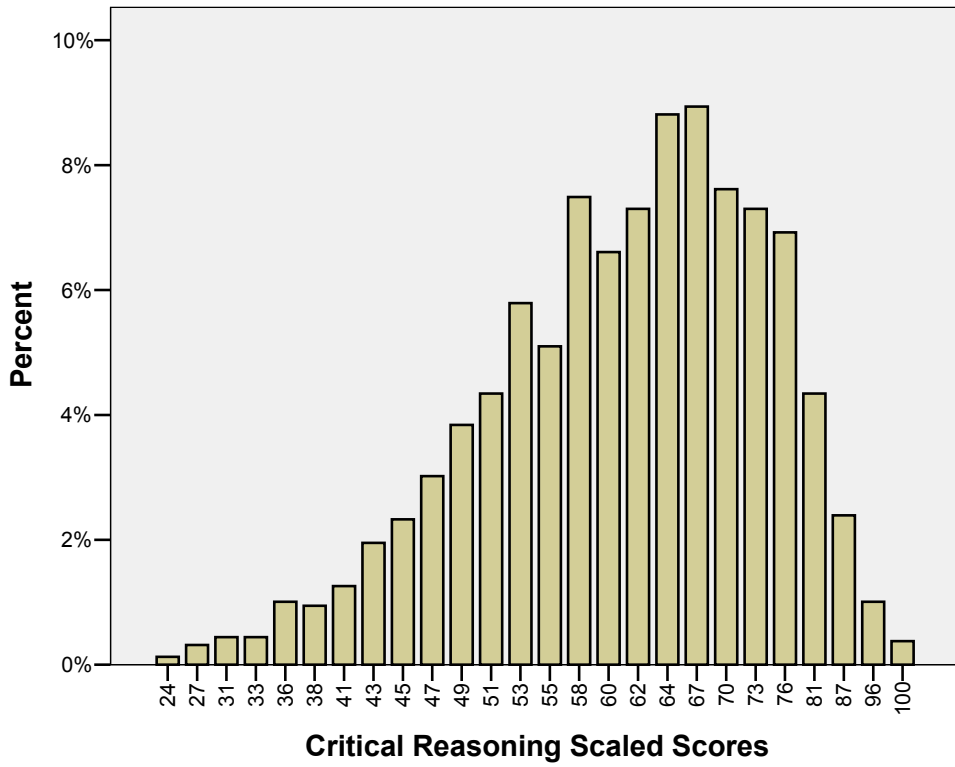
Joining the tops of the bars of each 2006 distribution gives a *unimodal* curve skewed to the right, indicating that many of the 2006 candidates score toward the upper end of the scale. (The bar charts for 2005 have curves that are more symmetrical or are skewed the other way.)

In all cases, the students are well spread by the test components and the total test according to performance.

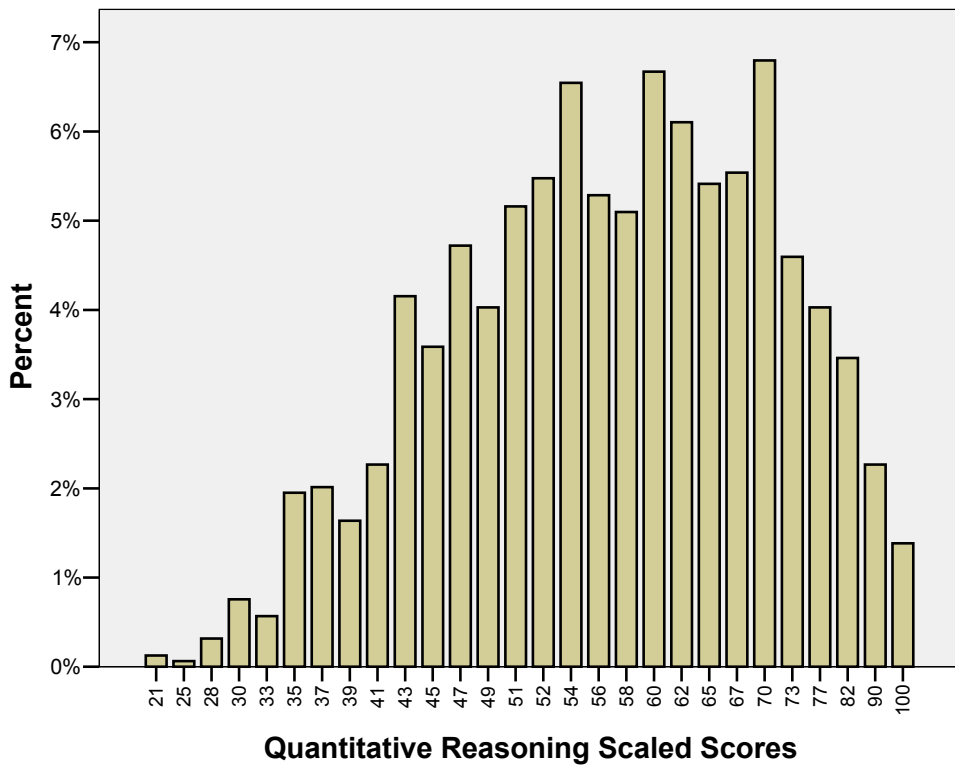
**Figure 3: Verbal-Plausible Scaled Scores**



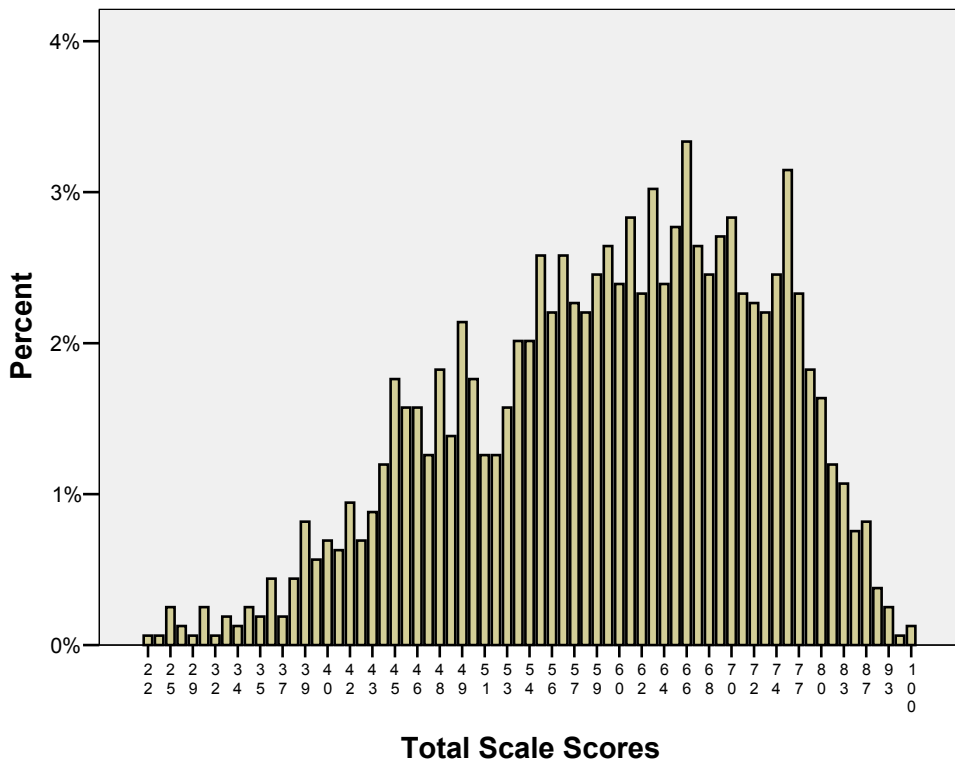
**Figure 4: Critical Reasoning Scaled Scores**



**Figure 5: Quantitative Reasoning Scaled Scores**

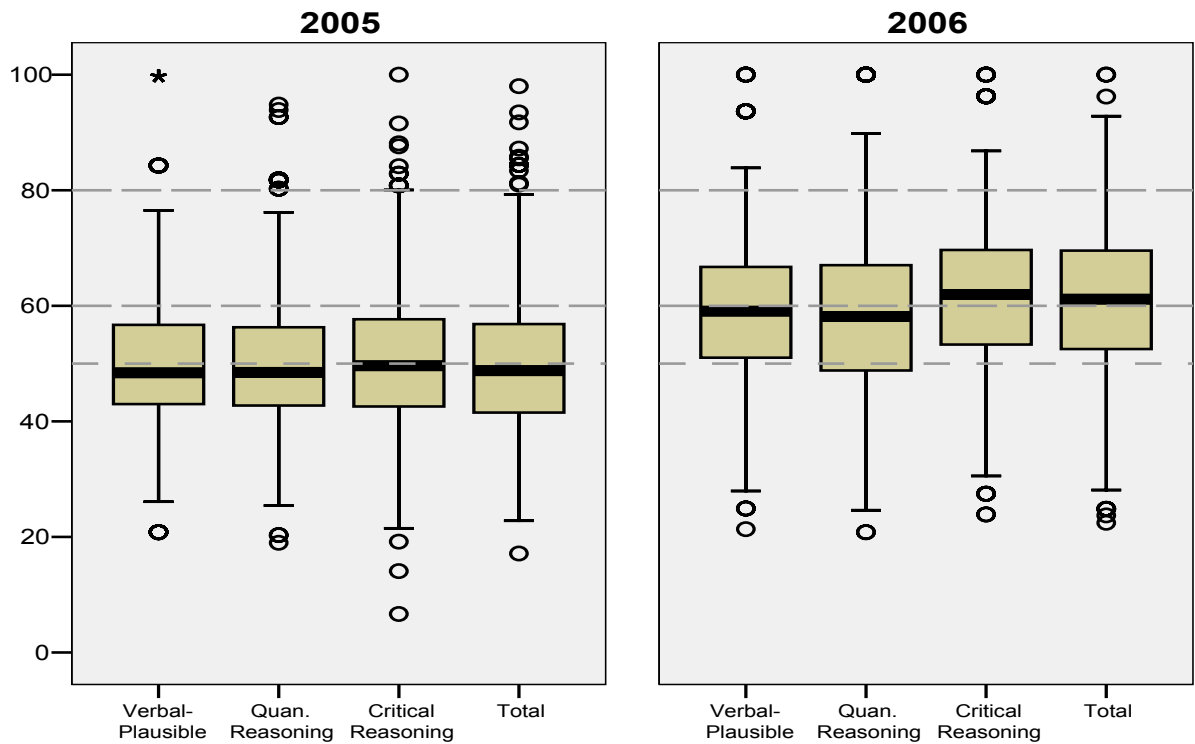


**Figure 6: Total Scale Scores**



Box-plots of the scaled scores below again show that the performance of the 2006 candidates exceeded that of the 2005 candidates.

**Figure 7: Box-plots of scaled scores for performance in 2005 and 2006**



## Analysis of the background of the sample

This section of the report contains summary statistics and analyses of the background of the sample of students participating in the study. Factors that can influence students' performance in the test (such as family or neighbourhood backgrounds) are investigated.

**Table 10: uniTEST scores. Summary statistics by gender**

| Gender | N    | Quantitative Reasoning Score |       | Critical Reasoning Score |       | Verbal-plausible Score |       | Total Score |       |
|--------|------|------------------------------|-------|--------------------------|-------|------------------------|-------|-------------|-------|
|        |      | Mean                         | SD    | Mean                     | SD    | Mean                   | SD    | Mean        | SD    |
| Male   | 652  | 61.7                         | 14.03 | 63.3                     | 12.34 | 58.7                   | 12.86 | 62.2        | 12.55 |
| Female | 937  | 56.5                         | 12.81 | 61.5                     | 12.04 | 58.9                   | 12.52 | 59.6        | 11.87 |
| All    | 1589 | 58.6                         | 13.57 | 62.2                     | 12.19 | 58.8                   | 12.66 | 60.7        | 12.21 |

Differences in test scores between males and females are statistically significant for the quantitative reasoning component, the critical reasoning component and the total score. However, it should be noted that although these differences are statistically significant, the overall difference is only around one fifth of a Standard Deviation.

For mean GCSE, gender differences are not significant (mean GCSE for males is 7.0019 and for females is 7.0586).

This differential gender performance will be discussed in greater detail later in this report.

The following table (Table 11) shows some summary statistics of the uniTEST scores by attainment group (mean GCSE). These groups were set to have around the same numbers of students. In all three components of the test, and in the test as a whole, students with higher prior achievement obtain higher scores. Differences in test scores between attainment groups are statistically significant. There is, however, some overlap between groups.

**Table 11: uniTEST scores. Summary statistics by attainment group**

| Attainment Group<br>(mean GCSE) | Quantitative Reasoning Score |       | Critical Reasoning Score |       | Verbal-plausible Score |       | Total Score |       |
|---------------------------------|------------------------------|-------|--------------------------|-------|------------------------|-------|-------------|-------|
|                                 | Mean                         | SD    | Mean                     | SD    | Mean                   | SD    | Mean        | SD    |
| Group I (0 – 6.5)               | 48.9                         | 10.84 | 53.7                     | 11.08 | 51.4                   | 10.41 | 51.1        | 10.29 |
| Group II (6.6 – 7.0)            | 56.0                         | 11.50 | 60.8                     | 9.53  | 56.7                   | 10.23 | 58.4        | 9.15  |
| Group III (7.1 – 7.4)           | 60.2                         | 12.64 | 63.8                     | 10.52 | 59.8                   | 10.36 | 62.2        | 9.98  |
| Group IV (7.5 – 7.7)            | 64.4                         | 12.46 | 68.1                     | 9.95  | 63.6                   | 11.20 | 66.7        | 9.33  |
| Group V (7.8 – 8.0)             | 68.2                         | 11.69 | 70.8                     | 9.58  | 68.0                   | 10.60 | 71.0        | 9.22  |

There is some evidence from these results that the uniTEST reveals academic potential in some candidates that, for whatever reason, have not done particularly well in their GCSEs. It seems reasonable to assume that, for some candidates, their reasoning abilities are not reflected in their GCSE results.

Students' choices of degree and institution are usually related to their prior attainment. In the following table we display the relationship between students' first choice of institutions and prior attainment.

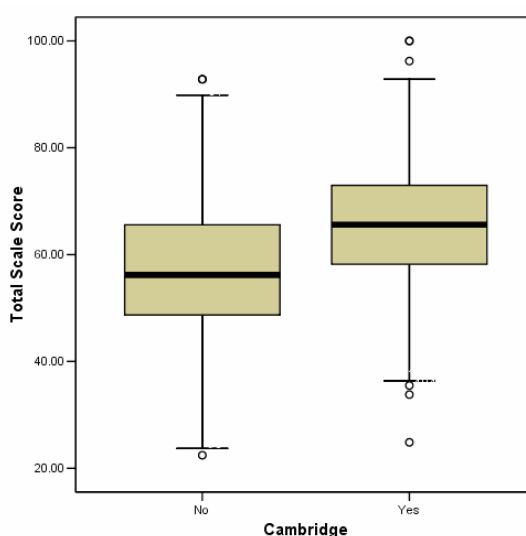
**Table 12: Institution and ability group (number of students)**

| University                  | Attainment Group |          |           |          |                 |
|-----------------------------|------------------|----------|-----------|----------|-----------------|
|                             | low<br>Group I   | Group II | Group III | Group IV | high<br>Group V |
| University of Bristol       | 26               | 39       | 35        | 29       | 25              |
| University of Cambridge     | 23               | 61       | 106       | 127      | 155             |
| University of Durham        | 32               | 30       | 20        | 21       | 12              |
| University of Edinburgh     | 22               | 15       | 14        | 5        | 4               |
| University of Exeter        | 59               | 25       | 19        | 12       | 3               |
| University of Hertfordshire | 36               | 5        | 0         | 0        | 0               |
| University of Warwick       | 50               | 48       | 39        | 32       | 15              |

As expected, students with high mean GCSE aspire to go to Cambridge (72% of the students in the highest attainment group had applied to Cambridge) or “Premier League” universities (ranked 3 to 10).

Although uniTEST is designed for use by a broad range of HE institutions it is, nevertheless, able to distinguish a wide range of ability levels amongst applicants to even the most selective institutions. The uniTEST scores of students applying to Cambridge appear to be well spread. This is illustrated in Figure 8.

**Figure 8: Test scores and Cambridge applications.**



The students who took the test have applied to 64 different courses in seven participating universities. For ease of analysis, students have been assigned to Field of Study groups, based on the degree courses to which they have applied.

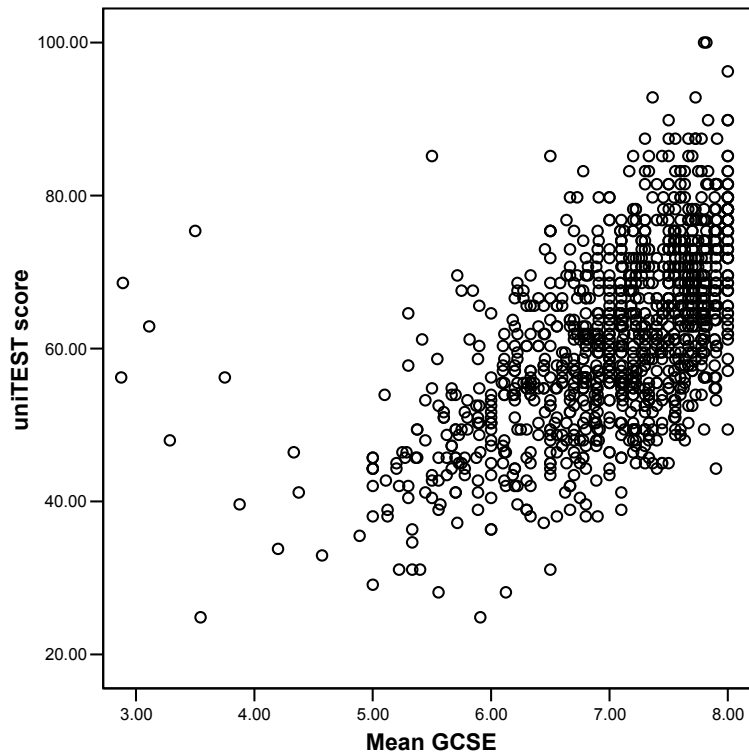
**Table 13: uniTEST scores. Summary statistics by field of study**

| Field of study           | N   | Quantitative Reasoning Score |       | Critical Reasoning Score |       | Verbal-Plausible Score |       | Total Score |       |
|--------------------------|-----|------------------------------|-------|--------------------------|-------|------------------------|-------|-------------|-------|
|                          |     | Mean                         | SD    | Mean                     | SD    | Mean                   | SD    | Mean        | SD    |
| Arts and Humanities      | 496 | 55.3                         | 11.67 | 63.4                     | 11.61 | 62.2                   | 11.31 | 61.1        | 10.92 |
| Business/Commerce        | 202 | 60.6                         | 13.16 | 60.8                     | 11.84 | 55.3                   | 12.22 | 59.7        | 12.09 |
| Computers/IT             | 16  | 64.3                         | 11.42 | 65.4                     | 11.99 | 60.0                   | 11.27 | 64.7        | 11.18 |
| Engineering/Architecture | 79  | 66.0                         | 12.21 | 64.3                     | 10.94 | 57.8                   | 11.66 | 64.2        | 11.62 |
| Education/Social         | 11  | 46.1                         | 13.74 | 49.2                     | 11.08 | 49.3                   | 12.35 | 47.7        | 12.24 |
| Law/Legal                | 135 | 55.9                         | 13.29 | 60.9                     | 12.76 | 59.7                   | 14.36 | 59.4        | 13.03 |
| Medicine/Dentistry       | 117 | 61.4                         | 11.31 | 65.0                     | 10.84 | 60.5                   | 12.17 | 63.4        | 10.40 |
| Nursing                  | 14  | 45.5                         | 7.35  | 54.3                     | 9.77  | 51.5                   | 11.74 | 50.0        | 9.16  |
| Science/Maths            | 519 | 62.8                         | 14.73 | 64.2                     | 11.90 | 58.9                   | 11.50 | 63.0        | 12.24 |

Differences in test scores between all fields of study are statistically significant. The students who obtained the highest scores in the test are those applying for degrees in the field of Engineering/Architecture and Computers/IT. The lowest scores are obtained by students applying for degrees in the field of Education/Social and Nursing. (It must be noted that these differences may be affected by the dominance of certain subjects by applicants to certain universities.)

One interesting thing to explore is whether uniTEST scores are related to students' GCSE results and, if so, the strength of the relationship. Figure 9 shows the relationship between the uniTEST scores and the mean GCSE for the candidates with matched records.

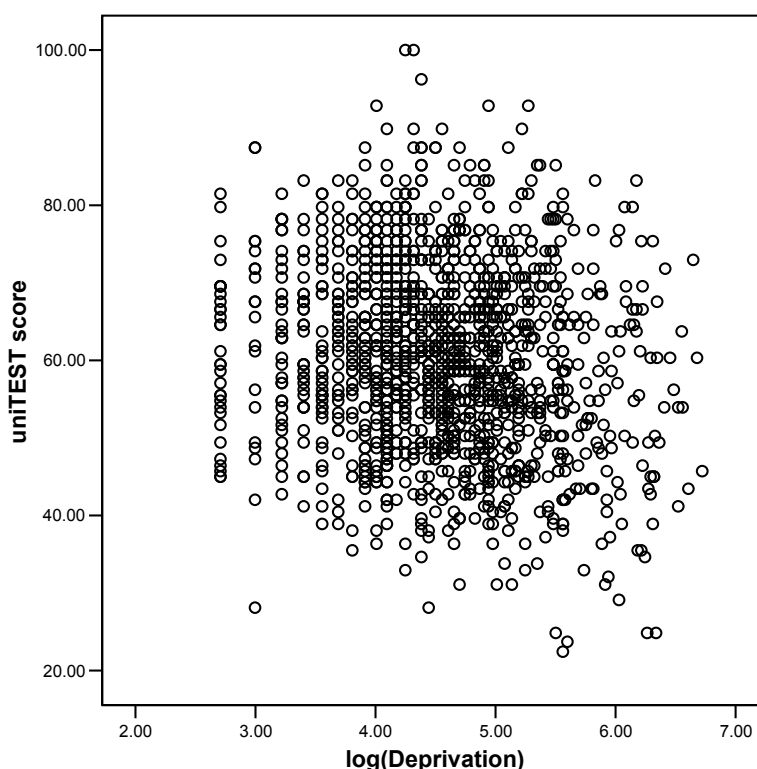
**Figure 9: uniTEST scores and GCSE scores**



Although the uniTEST scores are clearly related to mean GCSE (correlation coefficient 0.56), there are students who have high scores in the test but who are not in the top GCSE attainment group. For example, there are students with a grade D (on average) who are above the average mark in the uniTEST. These are students who may have the potential to do well in Higher Education but whose GCSE scores do not demonstrate this. (In the 2005 study, the correlation between test scores and mean GCSE was 0.64, perhaps because of the broader range of candidate ability in the sample.)

It is well known that students from socially deprived areas have, on average, lower educational attainment than their counterparts from more advantaged areas. Figure 10 shows uniTEST scores and deprivation index for all students who took uniTEST (a log transformation has been applied to this index).

**Figure 10: Deprivation and uniTEST scores (all candidates).**



There are some students from deprived areas who achieve above average scores in uniTEST. However, the highest scores tend to be achieved by people from less deprived areas (correlation coefficient -0.15).

### ***MLwiN modelling***

Neighbourhood characteristics and family background variables were considered using area information from census data, matched with students' home postcodes, and application data provided by UCAS.

Background factors that were considered include:

- deprivation,
- number of people in the area that have level 4/5 qualifications,
- percentage of pupils in the area achieving 5 or more GCSE A\*-C,
- distance travelled to work,
- number of lone parent households with dependent children,
- institution students applied to,
- field of study which students applied to.

To check on background effects in uniTEST, regression models were fitted (multilevel models with candidates nested into schools). For each, the score of the test is the dependent variable and background factors are introduced as independent variables. Student and school characteristics such as disabilities, ethnicity, socio-economic status and school type were also included in the models.

After adjusting for GCSE performance, only a few of the variables were significant. The effects of gender, mean GCSE, ethnicity and those background variables found to have a significant effect on test scores are shown in Figure 11. Any variable whose line intersects with the vertical zero axis can be regarded as not significant (at the 5%

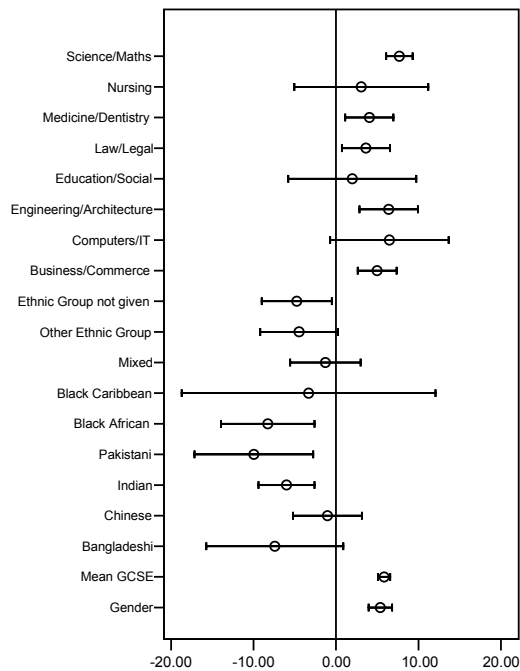
level). Positive values imply a positive relationship with the outcome; negative values imply that the uniTEST score decreases with higher values of the background variable. From Figure 11, it can be seen at a glance which variables are strongly related to the uniTEST score, both positively and negatively, and which ones seem to have much less definite relationships, even if they are statistically significant.

Deprivation, number of people in the area that have level 4/5 qualifications, percentage of pupils in the area achieving 5 or more GCSE A\*-C, distance travelled to work and number of lone parent households with dependent children do not seem to be associated with the students' success in the uniTEST. The effects of centre type, students' disabilities and socio-economic group were also not significant and therefore are not presented here.

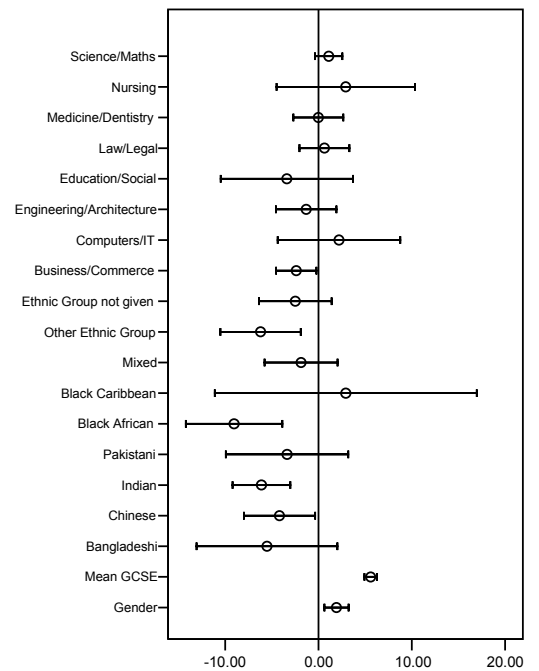
**Figure 11: Effects of students' characteristics on the uniTEST**

*The reference category for field of study is Arts/Humanities and for ethnicity group white.*

*Quantitative Reasoning*



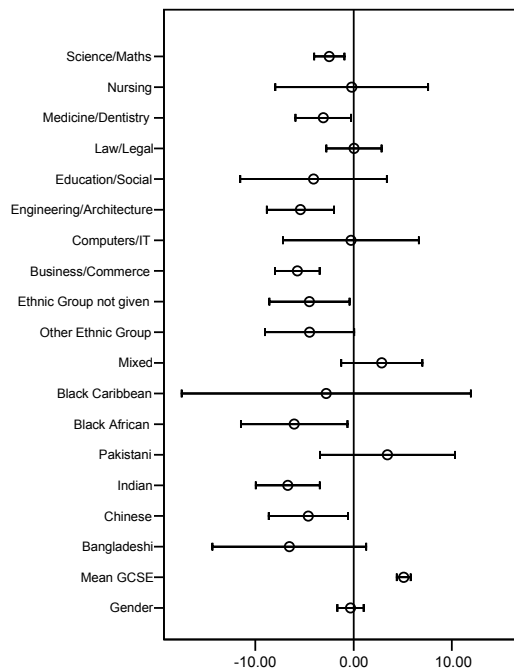
*Critical Reasoning*



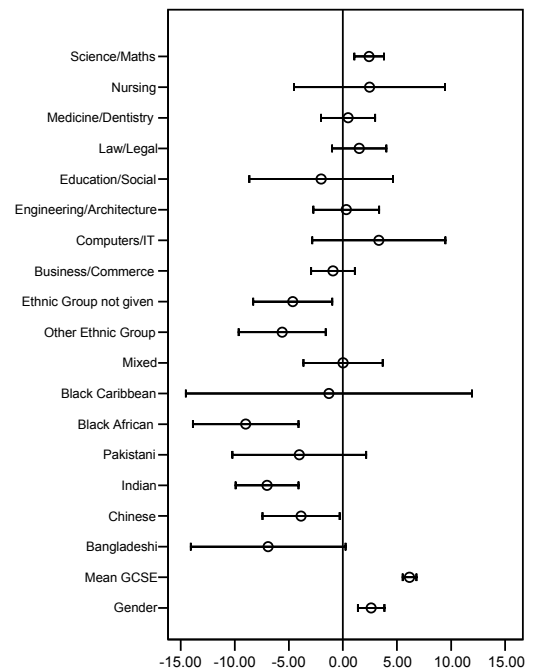
Continued overleaf

**Figure 11 continued**

*Verbal-Plausible Reasoning*



*Total uniTEST Score*



For example, *once prior attainment at GCSE is allowed for*, Science/Maths applicants perform significantly better in Quantitative Reasoning, and less well in Verbal-Plausible Reasoning than Arts/Humanities applicants. Although, *after allowing for performance at GCSE*, there is some indication that applicants for Nursing perform slightly better than Arts/Humanities applicants on some elements of the test, these differences are not statistically significant.

The variable that has the largest positive effect on uniTEST total score is prior attainment (mean GCSE).

Substantial differences appeared between ethnic groups. The results show that in comparison to the 'White' group, other ethnic groups such as Bangladeshi, Black African, Chinese or Indian, generally did not perform as well in uniTEST overall, although comparisons of performance on different test components reveal some intriguing patterns.

Figure 11 indicates that, although the prior attainment has the highest impact on uniTEST score, other factors such as gender or ethnicity explain a proportion of the variation in students' outcomes.

## Person-Item Analysis

This section examines the psychometric properties of the multiple-choice items.

Standard Item Response Theory (IRT) techniques were employed in the analysis of the results<sup>1</sup>. For each component, QUEST Short Item Statistics and Person-Item Plots are presented.

### Item Statistics

The item statistics given for each numbered item:

- *Facility*, which is the percentage of students answering correctly
- *Point Biserial* (Pt Bis), which is an index of the item's ability to discriminate between more and less able students (items with a point biserial below 0.19 would rarely be used).
- *$\alpha$ -deleted*, which indicates what happens to test reliability if the item is removed
- *Diff*, which is an index of item difficulty in logits. (The higher the facility, the lower the logit value. The logits are assigned in such a way that the average of the item logit values for an analysis is equal to zero.)
- *Infit MNSQ*, which is an index of how closely an item's performance fits<sup>2</sup> with the performance of the other items measuring the single construct dimension. In accordance with the Rasch Item Response Theory model<sup>3</sup>, items with relatively high INFIT MNSQ values (certainly those above 1.20) should be examined to see if they deviate substantively from the others in the set intended to measure a singular, coherent dimension.

Also provided is an index of test/component *Reliability*, in this case *Internal Consistency*, which is expected to be around 0.8 .

### Person-Item Plots

Another graphical output from QUEST is the Person-Item plot. This presents on the same vertical logit scale as the item difficulty estimates and student ability estimates. Students and items at the higher end are understood to be, respectively, more able or more difficult.

Items in the same units are grouped in the same vertical arrays.

By convention, the average of the logit values for the items for an analysis is set to zero to define the origin of the scale.

---

<sup>1</sup> For the multiple-choice items, which were scored dichotomously, a Rasch model was used to estimate candidate ability and item difficulty. See Adams, R.J. & Khoo, S-K. (1994). *Quest: the Interactive Test Analysis System*. Camberwell, Vic, Australia: Australian Council for Educational Research.

<sup>2</sup> To measure how well the items worked together to represent a single underlying trait, a fit statistic is generated for each question. This statistic represents the difference between the modelled response and the observed responses. The weighted mean square fit statistic (INFIT MNSQ) provides a measure of the item coherence to the underlying trait.

<sup>3</sup> Rasch, G. (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Education Research, 1960 (expanded edition, Chicago: The University of Chicago Press).

When a candidate's ability is located at the same point as a particular item difficulty on such a scale, the Rasch model predicts that the candidate would have a 50% chance of correctly answering that item. It is at this point that the item is supplying the maximum amount of information about a candidate's ability and therefore maximising the precision of the measure of the ability that is reported.

The more items located in a region of a scale, the more precise the measure of ability made in this region. If high discrimination is required in another area of the scale, a modified instrument should be considered.

On a logit scale, the same difference in logits always represents the same difference in ability/difficulty.

## **Quantitative Reasoning**

As shown, the item facilities are well spread, stretching from 26.7 to 91.8, with a mean facility of 65.17. The mean test score is 19.39 out of 30, with a standard deviation of 5.76.

The reliability (internal consistency) of the QR component is 0.86, which is the highest of the three scales. This means that only about 14% of the variation in the observed scores is due to error of measurement, which is relatively low for a 30-item sub-test. No item has a point biserial below 0.30, with an average value of 0.43, which explains the high reliability.

The  $\alpha$ -deleted column indicates that the removal of any one item (except item 55) from the scale will lead to a drop in the reliability below 0.857. However, dropping item 55 makes no difference to reliability. This means that every other item is needed to maintain the high reliability achieved.

As indicated in the table and the Person-Item Plot for QR, the difficulty estimates of the 30 items range from -1.95 to 2.24, spanning a range of 4.19 logits. However, this is not as wide as the ability range of the candidates which stretches from -2 logits to 4 logits. The measured ability range of the candidates is extremely broad, with a group of very able candidates at the top end.

The mean ability estimate of the 2006 candidates, who are much stronger on average than the 2005 group, is 0.93 logit above the average difficulty of the items. Although the test matches well the 2005 group, for groups of the high ability of the 2006 group some more difficult items, preferably above 2.5 logits, would be useful.

Most of the items have Infit Mean Squares not exceeding 1.10. There are two items with relatively large values (item 41 with 1.19 and item 55 with 1.21), which may focus on skills different from the other items.

**Table 14: Quest Item Analysis Results for Quantitative Reasoning**

uniTEST 2006 - missing as missing: all on QR (N = 1589 L = 30 Probability Level=0.50)

| Item | Facility | Pt Bis | $\alpha$ -deleted | Diff  | Infit MNSQ |
|------|----------|--------|-------------------|-------|------------|
| 6    | 91.0     | 0.35   | .855              | -1.84 | 0.94       |
| 7    | 83.8     | 0.39   | .854              | -1.09 | 0.98       |
| 9    | 89.1     | 0.39   | .854              | -1.60 | 0.94       |
| 10   | 58.6     | 0.50   | .851              | 0.48  | 0.99       |
| 25   | 90.6     | 0.35   | .855              | -1.78 | 0.94       |
| 26   | 44.7     | 0.58   | .849              | 1.20  | 0.86       |
| 27   | 52.8     | 0.59   | .848              | 0.79  | 0.86       |
| 34   | 86.6     | 0.45   | .852              | -1.34 | 0.89       |
| 35   | 75.3     | 0.44   | .853              | -0.46 | 0.99       |
| 40   | 76.7     | 0.50   | .851              | -0.55 | 0.91       |
| 41   | 55.9     | 0.35   | .856              | 0.63  | 1.19       |
| 42   | 91.8     | 0.32   | .855              | -1.95 | 0.95       |
| 43   | 84.5     | 0.38   | .854              | -1.15 | 0.98       |
| 44   | 65.2     | 0.51   | .850              | 0.14  | 0.95       |
| 51   | 81.2     | 0.42   | .853              | -0.87 | 0.97       |
| 54   | 89.7     | 0.32   | .855              | -1.67 | 1.00       |
| 55   | 63.5     | 0.31   | .857              | 0.23  | 1.21       |
| 56   | 55.9     | 0.43   | .853              | 0.63  | 1.08       |
| 62   | 68.6     | 0.38   | .854              | -0.06 | 1.10       |
| 63   | 31.8     | 0.38   | .855              | 1.92  | 1.09       |
| 64   | 62.0     | 0.40   | .854              | 0.31  | 1.10       |
| 65   | 26.7     | 0.47   | .852              | 2.24  | 0.95       |
| 73   | 73.0     | 0.56   | .849              | -0.32 | 0.86       |
| 74   | 83.1     | 0.46   | .852              | -1.02 | 0.92       |
| 75   | 49.1     | 0.47   | .852              | 0.98  | 1.03       |
| 76   | 41.4     | 0.56   | .849              | 1.38  | 0.88       |
| 77   | 50.1     | 0.42   | .854              | 0.93  | 1.09       |
| 85   | 55.1     | 0.47   | .852              | 0.68  | 1.02       |
| 89   | 50.0     | 0.50   | .851              | 0.95  | 0.98       |
| 90   | 27.3     | 0.39   | .854              | 2.21  | 1.04       |

Mean test score 19.39  
 Standard deviation 5.76  
 Internal Consistency 0.857

**General Facility Statistics**

| N           | Valid   | 30    |
|-------------|---------|-------|
|             | Missing | 0     |
| Mean        |         | 65.17 |
| Range       |         | 65.10 |
| Minimum     |         | 26.70 |
| Maximum     |         | 91.80 |
| Percentiles | 25      | 50.08 |
|             | 50      | 64.35 |
|             | 75      | 83.98 |

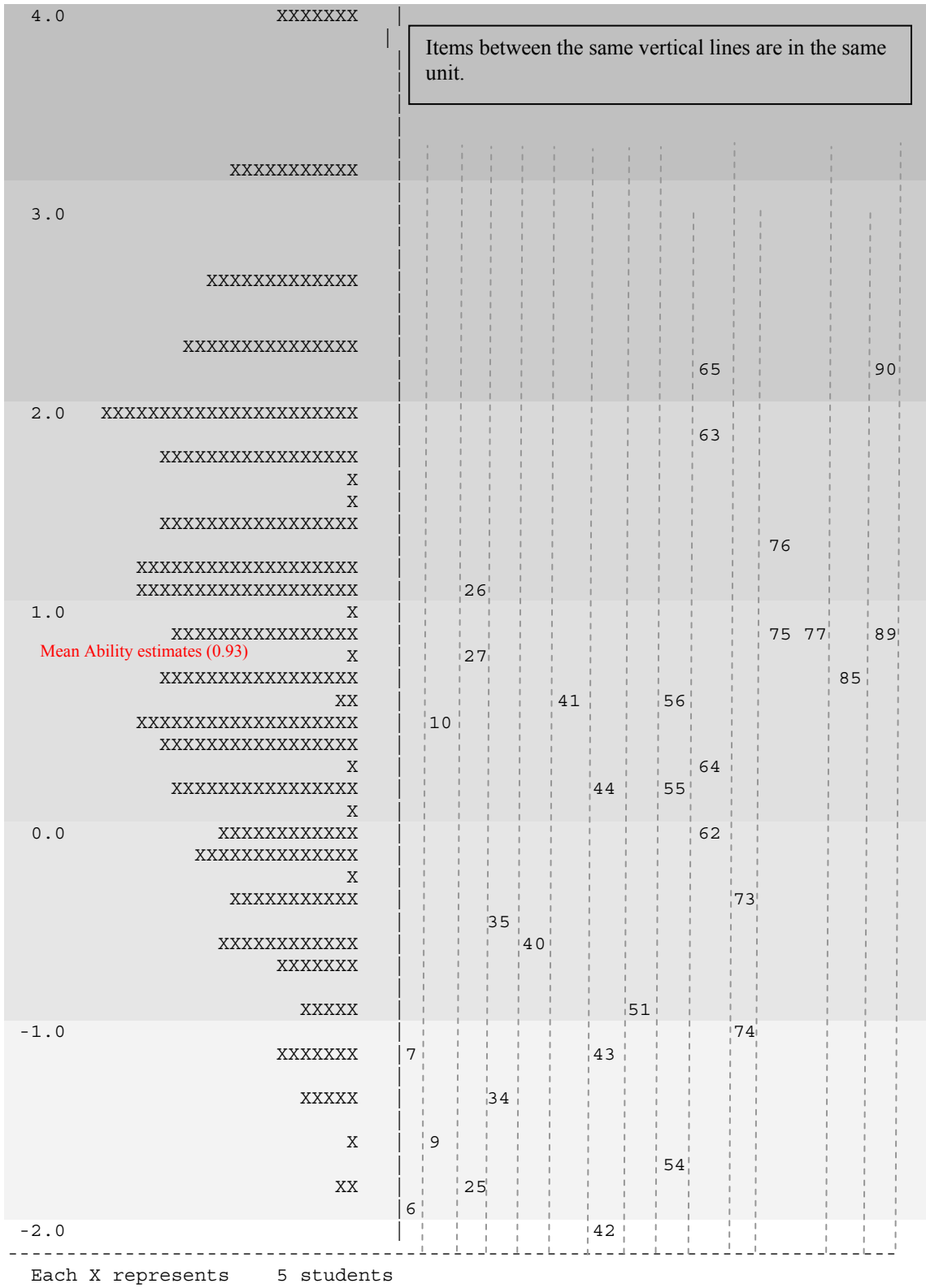
### Figure 12: Person-Item Plot Quantitative Reasoning

uniTEST 2006 - missing as missing

Item Estimates (Thresholds)

27/ 6/2006 10:51

all on QR (N = 1589 L = 30 Probability Level=0.50)



## Critical Reasoning

As shown, the item facilities stretch from 39.9 to 98.3, with a mean of 68.85. The mean test score is 20.52 with a standard deviation of 4.77.

On the whole, items in this component have lower point biserial values than items in the other two components, the CR component having a reliability (internal consistency) of 0.772. The point biserials have an average value of 0.36. Given that the reliability is still close to 0.8, it is satisfactory. Because the CR component has three sub-components, it is more diverse than the other two components, which helps explain its lower internal consistency. (The 30 CR items are divided into groups of ten, one for each sub-component tapping on a different type of critical reasoning, labelled in the item-person map as CR1, CR2 and CR3.) It may be that the sub-components could be used differentially for different course selection decisions.

The  $\alpha$ -deleted column indicates that the removal of any one item, other than item 67, from the scale will lead to a drop in the reliability. Removal of item 67 from the scale leads to a slight increase of reliability 0.772 to 0.775.

As indicated in the table and the Person-Item Plot, the difficulty estimates of the 30 items range from -3.25 to 1.55, spanning a wide range of 4.8 logits. The mean person ability estimate is 1.09 logit above the average difficulty of the items. Although the test matches well the 2005 group, for groups of the high ability of the 2006 group some more difficult items, preferably above 2.5 logits, would be useful.

Most of the items have Infit Mean Squares not exceeding 1.10. There is one item with a relatively large value (item 67), which may focus on skills different from the other items.

**Table 15: Item Analysis Results for Critical Reasoning**

uniTEST 2006 - missing as missing: all on CR (N = 1589 L = 30 Probability Level=0.50)

| Item | Facility | Pt Bis | $\alpha$ -deleted | Diff  | Infit MNSQ |
|------|----------|--------|-------------------|-------|------------|
| 8    | 98.3     | 0.19   | .771              | -3.25 | 0.97       |
| 11   | 75.3     | 0.48   | .760              | -0.19 | 0.90       |
| 12   | 70.2     | 0.37   | .765              | 0.10  | 1.00       |
| 13   | 90.3     | 0.22   | .770              | -1.39 | 1.04       |
| 15   | 95.2     | 0.29   | .768              | -2.19 | 0.95       |
| 16   | 88.3     | 0.37   | .765              | -1.18 | 0.93       |
| 17   | 89.7     | 0.37   | .765              | -1.33 | 0.92       |
| 18   | 79.1     | 0.47   | .760              | -0.43 | 0.89       |
| 19   | 62.0     | 0.43   | .762              | 0.51  | 0.97       |
| 28   | 68.9     | 0.30   | .770              | 0.16  | 1.08       |
| 29   | 65.8     | 0.42   | .763              | 0.33  | 0.97       |
| 30   | 64.9     | 0.29   | .771              | 0.37  | 1.09       |
| 32   | 80.1     | 0.38   | .764              | -0.50 | 0.97       |
| 33   | 77.0     | 0.41   | .763              | -0.30 | 0.96       |
| 49   | 73.1     | 0.50   | .758              | -0.06 | 0.88       |
| 50   | 66.5     | 0.47   | .760              | 0.29  | 0.92       |
| 52   | 58.1     | 0.31   | .769              | 0.70  | 1.08       |
| 53   | 71.8     | 0.44   | .761              | 0.01  | 0.94       |
| 66   | 55.2     | 0.31   | .769              | 0.84  | 1.08       |
| 67   | 48.8     | 0.20   | .775              | 1.14  | 1.18       |
| 68   | 76.0     | 0.29   | .768              | -0.23 | 1.06       |
| 69   | 57.7     | 0.32   | .768              | 0.72  | 1.07       |
| 70   | 82.7     | 0.45   | .760              | -0.68 | 0.89       |
| 71   | 73.3     | 0.35   | .766              | -0.07 | 1.01       |
| 72   | 48.8     | 0.40   | .764              | 1.13  | 1.00       |
| 83   | 45.7     | 0.40   | .764              | 1.28  | 0.98       |
| 84   | 39.9     | 0.44   | .761              | 1.55  | 0.93       |
| 86   | 50.8     | 0.32   | .768              | 1.05  | 1.06       |
| 87   | 63.0     | 0.26   | .770              | 0.48  | 1.11       |
| 88   | 49.0     | 0.31   | .769              | 1.13  | 1.07       |

Mean test score           20.52  
 Standard deviation       4.77  
 Internal Consistency     0.772

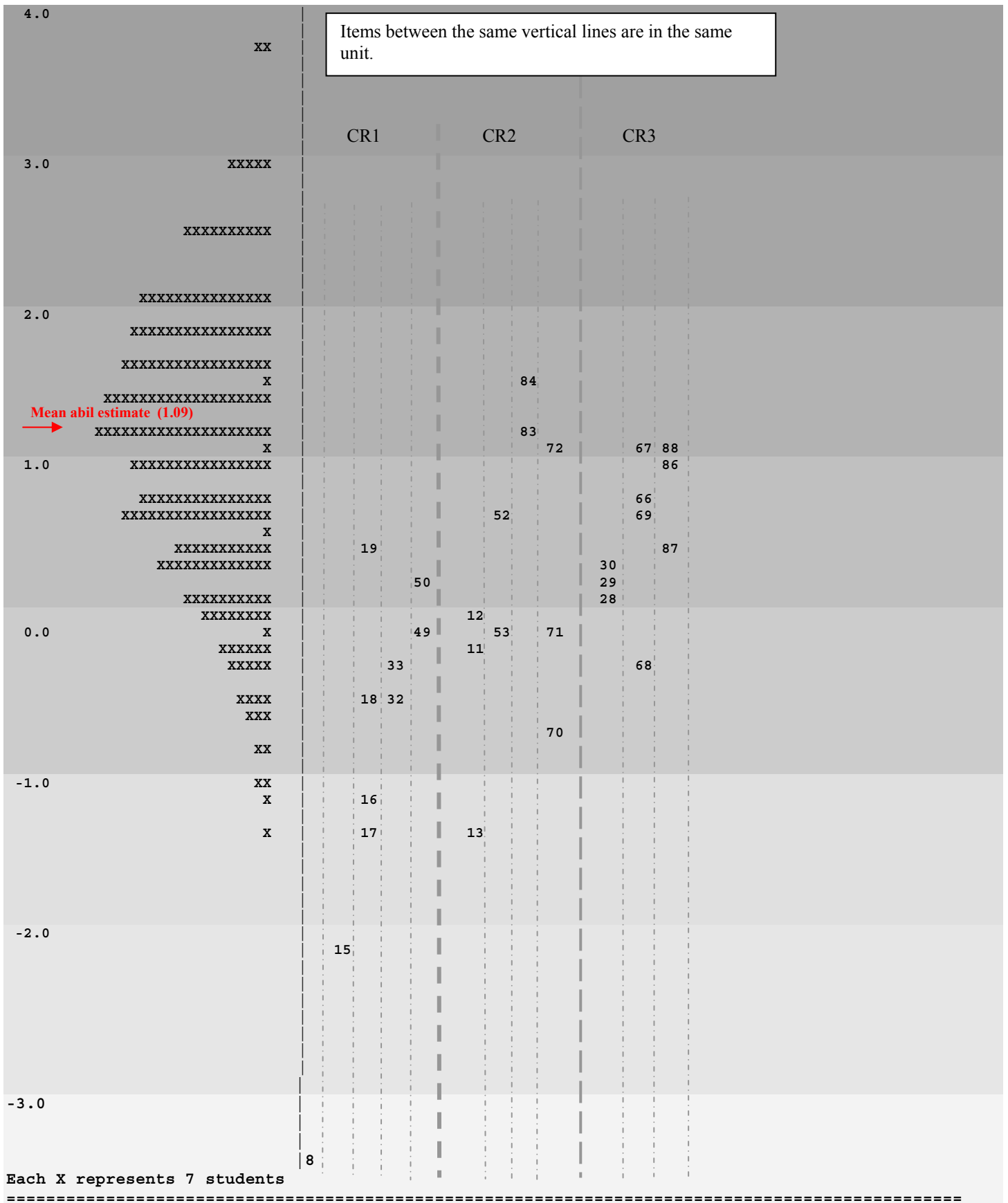
**General Facility Statistics**

| N           | Valid   | 30    |
|-------------|---------|-------|
|             | Missing | 0     |
| Mean        |         | 68.85 |
| Range       |         | 58.40 |
| Minimum     |         | 39.90 |
| Maximum     |         | 98.30 |
| Percentiles | 25      | 57.08 |
|             | 50      | 69.55 |
|             | 75      | 79.35 |

**Figure 13: Person-Item Plot Critical Reasoning**

uniTEST 2006 - missing as missing Scale for Critical Reasoning

Item Estimates (Thresholds) 27/ 6/2006 all on CR (N = 1589 L = 30 Probability Level=0.50)



## Verbal-Plausible Reasoning

As shown, the VP component had a mean facility of 67.3, with facilities ranging from 38.4 to 82.8. Mean test score was 20.07 with a standard deviation of 5.18.

The reliability of the VP scale is 0.79, which is close to the 0.8 standard.

The  $\alpha$ -deleted column indicates that removal of items 5, 14, 37, 38 and 39 individually would lead to a slight increase in reliability, suggesting these items measure something different from the other items. Items 37-39 form a single unit.

As indicated in the table and the Person-Item Plot, the ability range of the students is -1.5 logit to 3.5 logit while the range for items is about -1.1 to 1.3. The mean ability estimate of the students is 0.88 logit above the average difficulty of the items. Although the test matches well the 2005 group, for groups of the high ability of the 2006 group some more difficult items, preferably above 2.5 logits, would be useful.

The Infit Mean Squares of these 30 items range from 0.84 to 1.15, with no items with a value above 1.15 though a few are in the 1.10 to 1.15 range.

**Table 16: Item Analysis Results for Verbal-Plausible Reasoning**

uniTEST 2006 - missing as missing: all on VP (N = 1589 L = 30 Probability Level=0.50)  
 Summary Item Analysis all on VP (N = 1589 L = 30 Probability Level=0.50)

| Item | Facility | Pt Bis | $\alpha$ -deleted | Diff  | Infit | MNSQ |
|------|----------|--------|-------------------|-------|-------|------|
| 1    | 69.5     | 0.40   | .785              | -0.09 | 0.98  |      |
| 2    | 73.6     | 0.24   | .792              | -0.32 | 1.11  |      |
| 3    | 83.0     | 0.32   | .787              | -0.93 | 1.00  |      |
| 4    | 76.0     | 0.27   | .790              | -0.45 | 1.07  |      |
| 5    | 51.0     | 0.28   | .791              | 0.81  | 1.11  |      |
| 14   | 78.5     | 0.24   | .791              | -0.61 | 1.08  |      |
| 20   | 77.2     | 0.30   | .789              | -0.53 | 1.03  |      |
| 21   | 53.9     | 0.38   | .786              | 0.67  | 1.02  |      |
| 22   | 73.4     | 0.48   | .781              | -0.30 | 0.91  |      |
| 23   | 60.1     | 0.57   | .776              | 0.39  | 0.84  |      |
| 24   | 72.7     | 0.50   | .780              | -0.26 | 0.89  |      |
| 31   | 68.6     | 0.28   | .790              | -0.04 | 1.08  |      |
| 36   | 78.7     | 0.31   | .788              | -0.62 | 1.02  |      |
| 37   | 67.7     | 0.22   | .793              | 0.01  | 1.14  |      |
| 38   | 68.5     | 0.25   | .791              | -0.03 | 1.11  |      |
| 39   | 60.8     | 0.23   | .793              | 0.35  | 1.15  |      |
| 45   | 64.1     | 0.28   | .790              | 0.19  | 1.10  |      |
| 46   | 65.0     | 0.36   | .786              | 0.14  | 1.02  |      |
| 47   | 69.8     | 0.35   | .786              | -0.10 | 1.02  |      |
| 48   | 77.1     | 0.51   | .779              | -0.52 | 0.87  |      |
| 57   | 66.4     | 0.41   | .784              | 0.08  | 0.98  |      |
| 58   | 78.5     | 0.46   | .781              | -0.61 | 0.91  |      |
| 59   | 65.3     | 0.51   | .779              | 0.14  | 0.89  |      |
| 60   | 72.6     | 0.53   | .777              | -0.25 | 0.86  |      |
| 61   | 58.9     | 0.45   | .781              | 0.45  | 0.94  |      |
| 78   | 70.2     | 0.34   | .786              | -0.12 | 1.03  |      |
| 79   | 39.2     | 0.40   | .784              | 1.37  | 0.98  |      |
| 80   | 66.5     | 0.48   | .780              | 0.08  | 0.91  |      |
| 81   | 40.6     | 0.40   | .784              | 1.30  | 0.98  |      |
| 82   | 71.9     | 0.40   | .783              | -0.20 | 0.98  |      |

Mean test score           20.07  
 Standard deviation       5.18  
 Internal Consistency     0.791

**General Facility Statistics**

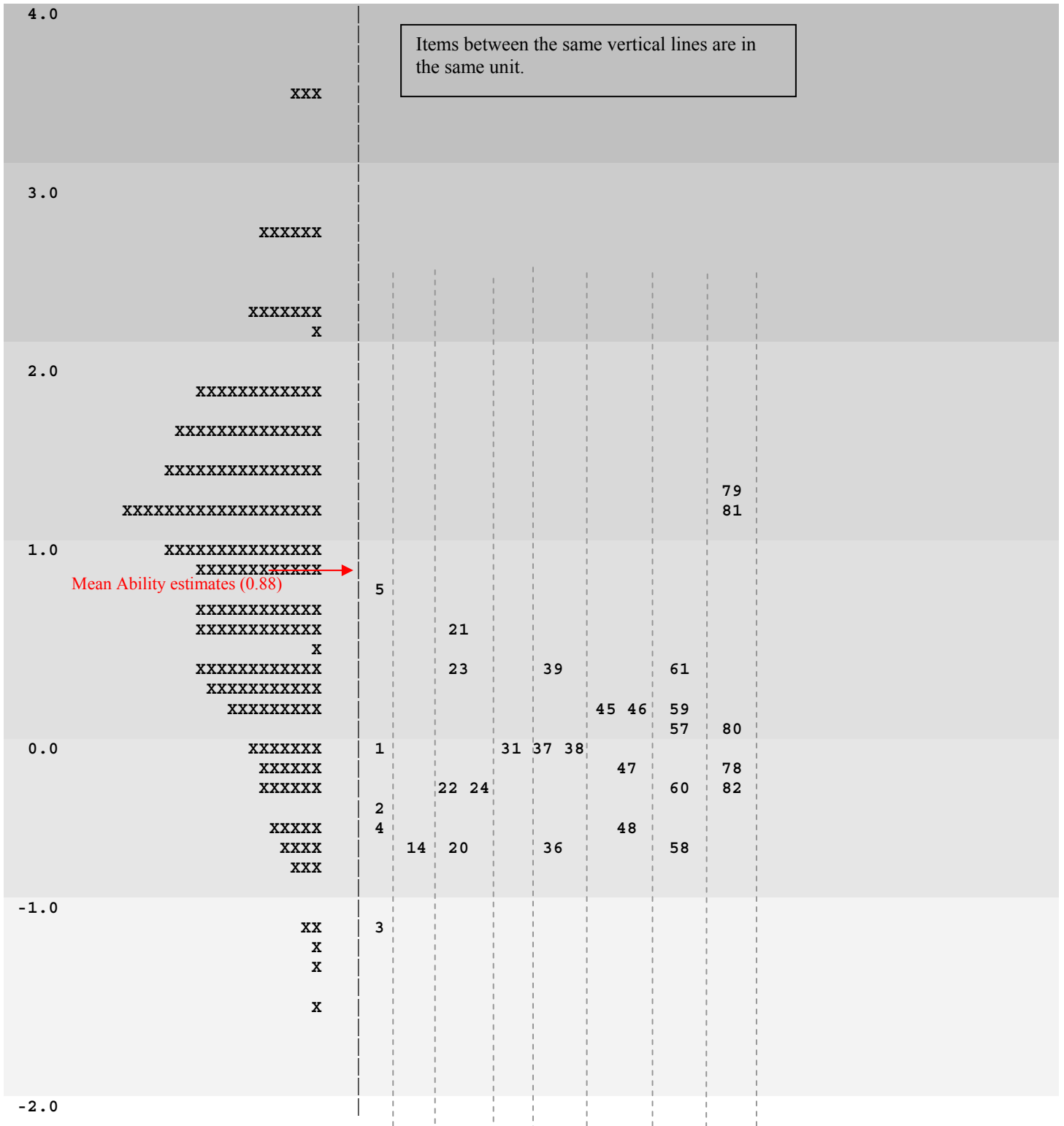
| N           | Valid   | 30    |
|-------------|---------|-------|
|             | Missing | 0     |
| Mean        |         | 67.31 |
| Range       |         | 43.80 |
| Minimum     |         | 39.20 |
| Maximum     |         | 83.00 |
| Percentiles | 25      | 63.28 |
|             | 50      | 69.05 |
|             | 75      | 74.20 |

**Figure 14: Person-Item Plot Verbal-Plausible Reasoning**

uniTEST 2006 - missing as missing

Item Estimates (Thresholds)  
all on VP (N = 1589 L = 30 Probability Level=0.50)

27/ 6/2006 10:51



Each X represents 8 students

## Differential Gender Performance

The following table shows that the mean scores on the four scales differ between the boys and girls.

The largest difference is found for QR, which has a difference of more than five points in favour of boys. On the other hand, girls did slightly better than boys on VP with a mean of 58.93 compared to 58.70 for boys.

**Table 17: Scaled Scores for Gender**

| Sex   |               | Total Scale Score | Verbal-Plausible Scale Score | Quantitative Reasoning Scale Score | Critical Reasoning Scale Score |
|-------|---------------|-------------------|------------------------------|------------------------------------|--------------------------------|
| Girls | Count         | 937               | 937                          | 937                                | 937                            |
|       | Mean          | 59.64             | 58.93                        | 56.47                              | 61.52                          |
|       | Std Deviation | 11.87             | 12.52                        | 12.81                              | 12.04                          |
|       | Maximum       | 92.82             | 100.00                       | 100.00                             | 100.00                         |
|       | Percentile 75 | 68.56             | 66.74                        | 64.59                              | 69.68                          |
|       | Median        | 59.48             | 59.09                        | 56.18                              | 62.01                          |
|       | Percentile 25 | 50.96             | 51.04                        | 46.96                              | 53.31                          |
|       | Minimum       | 24.85             | 24.92                        | 20.83                              | 23.90                          |
|       | Range         | 67.97             | 75.08                        | 79.17                              | 76.10                          |
| Boys  | Count         | 652               | 652                          | 652                                | 652                            |
|       | Mean          | 62.19             | 58.70                        | 61.69                              | 63.28                          |
|       | Std Deviation | 12.55             | 12.86                        | 14.03                              | 12.34                          |
|       | Maximum       | 100.00            | 100.00                       | 100.00                             | 100.00                         |
|       | Percentile 75 | 70.69             | 66.74                        | 69.82                              | 69.68                          |
|       | Median        | 62.88             | 59.09                        | 60.18                              | 64.31                          |
|       | Percentile 25 | 53.23             | 51.04                        | 52.50                              | 55.36                          |
|       | Minimum       | 22.44             | 21.35                        | 20.83                              | 23.90                          |
|       | Range         | 77.56             | 78.65                        | 79.17                              | 76.10                          |

Lavene's test of equality of variance shows that the data support the hypothesis of equal variances between boys and girls on all four scales with the exception of QR.

Independent samples *t*-tests indicate that the differences found between the means of boys and girls are significant at the 0.05 level for the total scale, the QR scale and the CR scale.

Therefore, the possibility of gender bias among the items was examined using the Mantel-Haenszel procedures which compare performance between different groups of boys and girls of the same ability using the total test scores as the matching criterion for ability.

The outcome of the procedures is a statistic called the Mantel-Haenszel common odds ratio. Following the practice of the ETS (Educational Testing Service), this statistic is transformed and then used to classify items into three categories referring to the degree of gender bias they exhibit.

“C” categorisation – represents moderate to large gender bias

“B” categorisation – represents slight to moderate gender bias

“A” categorisation – represents negligible gender bias

According to the ETS, items in the C category are subject to further scrutiny and may be eliminated from a test.

Analysis indicates for UniTEST 2006:

- No item falls in the “C” category.
- One VP item falls in the “B” category with boys slightly advantaged.
- One QR item falls in the “B” category with boys slightly advantaged.
- Four CR items fall in the “B” category, three of advantage to boys and one to girls.

Thus, the analysis indicates that gender bias is not a significant issue.

Differential performance on the basis of gender may be related to student interest and course participation.

If differential performance were to be a significant factor it could be addressed, for example, as follows: where students do not intend to proceed to quantitative courses, their QR score may not be used, or may be used in a minimal way to optimise correlation with course success (e.g. the QR component score could be ignored or be minimally weighted for non-quantitative Arts/Humanities courses). In this way, differential performance related to gender for the QR scale may be irrelevant.

## Factor Analysis

Factor analysis can be used to check on the dimensional structure of a test. It has been used here to check that uniTEST, at least, consists of three coherent and distinctive components.

## Exploratory Factor Analysis

The three components of uniTEST 2006 each consist of 30 items arranged in a total of 34 units - 8 VP units, 14 QR units and 12 CR units.

An *unrotated exploratory factor analysis* at unit level yields the following Scree plot which illustrates the presence of a dominant factor with an *eigenvalue* of almost 8.

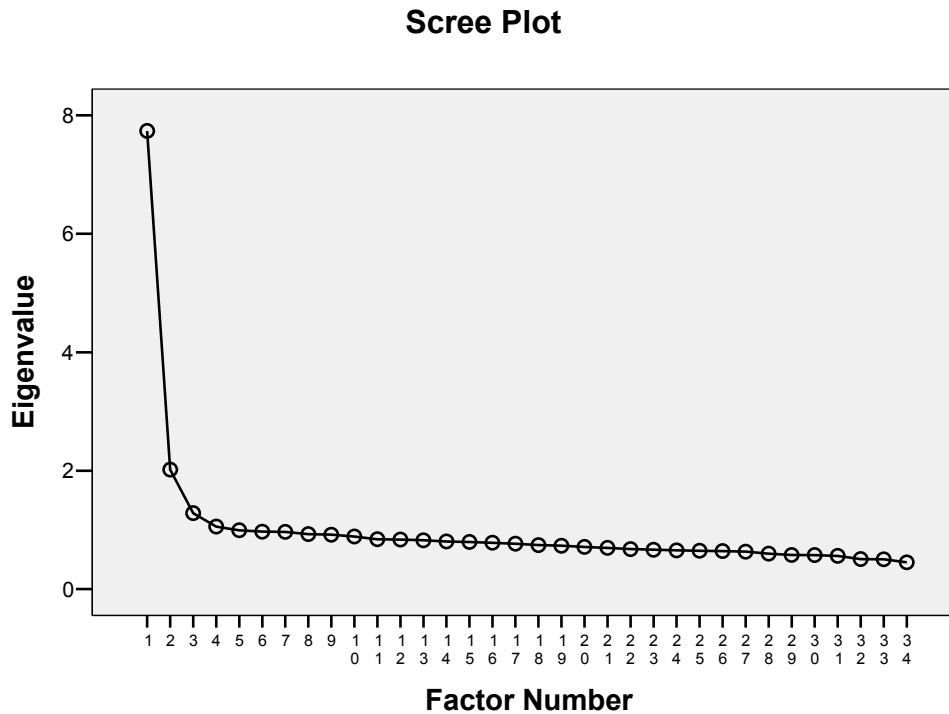
This justifies the use of a total scale as defined by all the 90 items together.

In fact, the *reliability of the total scale is 0.91*, leaving only about 9% of the variation in the total scaled score as error of measurement.

With the exception of the first three unrotated factors, all the other factor points are on a levelling curve running almost parallel to the x-axis, indicating that they are not significant.

The Scree plot thus suggests that a two or three factor solution could be a reasonable fit to the data, and this could be confirmed by rotated factor analysis.

**Figure 15: Scree plot**



The following table gives the rotated three factor loadings of the various units. Factor loadings that are very low are suppressed in the table.

Every unit has a loading of at least 0.1 on at least one factor dimension, and most have a loading of greater than 0.3 on one dimension.

Factor 1 is clearly associated with the QR units and Factor 2 with the VP units. There appears to be a third minor factor associated with some CR units, on which there is some loading from QR and VP units (units with a kind of deductive processing focus perhaps). However, most CR units tend to have loadings of considerable sizes on the other two factors. CR units that are text-based, interpretive and inductive tend to load on the VP factor whereas CR units that emphasise non-text based, deductive reasoning tend to load on the QR factor.

The CR items bridge the middle ground between QR and VP, thereby having value in the context of this test's likely use.

The factor loading observations are unsurprising, displaying the strong Q-V split seen in many tests, and agrees with the bivariate correlations presented earlier. (CR has a correlation of 0.662 with QR and a correlation of 0.654 with VP, while the correlation between VP and QR is just 0.499.)

**Table 18: Rotated Factor Matrix**

|       | Factor |      |      |
|-------|--------|------|------|
|       | 1      | 2    | 3    |
| u1VP  | .191   | .347 | .260 |
| u6VP  |        | .177 |      |
| u9VP  | .214   | .504 | .180 |
| u12VP |        | .227 |      |
| u15VP |        | .338 |      |
| u19VP | .185   | .503 |      |
| u24VP |        | .692 | .176 |
| u30VP | .272   | .584 |      |
| u2QR  | .368   |      | .265 |
| u4QR  | .428   | .174 | .268 |
| u10QR | .562   |      | .190 |
| u14QR | .501   |      | .220 |
| u16QR | .486   |      | .187 |
| u17QR | .256   |      |      |
| u18QR | .458   | .239 | .243 |
| u21QR | .415   |      | .175 |
| u23QR | .382   | .224 | .267 |
| u25QR | .447   | .312 |      |
| u28QR | .511   | .297 | .260 |
| u29QR | .594   | .272 |      |
| u32QR | .410   |      |      |
| u34QR | .489   | .205 |      |
| u3CR  |        |      | .396 |
| u7CR  |        |      | .482 |
| u8CR  | .398   | .227 | .343 |
| u13CR | .360   |      | .234 |
| u20CR | .467   | .253 | .216 |
| u5CR  | .231   | .468 | .212 |
| u22CR | .273   | .381 |      |
| u27CR | .425   | .337 |      |
| u31CR | .320   | .316 |      |
| u11CR |        | .389 |      |
| u26CR |        | .407 |      |
| u33CR |        | .417 |      |

Extraction Method: Alpha Factoring.  
 Rotation Equamax with Kaiser Normalization.

Thus, either a two or three factor solution may have a satisfactory fit to the data, with the CR units split if two factors are used.

Confirmatory factor analysis (next section) shows that a three factor solution aligned with the test components is a good enough fit to the data.

## Confirmatory Factor Analysis

A confirmatory factor analysis was performed using LISREL 8.5 using the unit assignment provided by the test developers.

**Table 19: Confirmatory Factor Analysis**

| Units | No of items | Name of unit              | Factor1 | Factor2 | Factor3 |
|-------|-------------|---------------------------|---------|---------|---------|
| u1VP  | 5           | <b>Paradox</b>            | 0.299   | 0       | 0       |
| u6VP  | 1           | <b>Police Cartoon</b>     | 0.041   | 0       | 0       |
| u9VP  | 5           | <b>National Character</b> | 0.473   | 0       | 0       |
| u12VP | 1           | <b>Dali</b>               | 0.051   | 0       | 0       |
| u15VP | 4           | <b>New York</b>           | 0.178   | 0       | 0       |
| u19VP | 4           | <b>A Life Apart</b>       | 0.339   | 0       | 0       |
| u24VP | 5           | <b>Eskimos</b>            | 0.548   | 0       | 0       |
| u30VP | 5           | <b>Social Contract</b>    | 0.485   | 0       | 0       |
| u2QR  | 2           | <b>Rectangles 3</b>       | 0       | 0.143   | 0       |
| u4QR  | 2           | <b>Road Offences</b>      | 0       | 0.224   | 0       |
| u10QR | 3           | <b>Income Tax</b>         | 0       | 0.390   | 0       |
| u14QR | 2           | <b>Symbol Fractions</b>   | 0       | 0.210   | 0       |
| u16QR | 1           | <b>Sea Shells</b>         | 0       | 0.129   | 0       |
| u17QR | 1           | <b>Alf and Bill</b>       | 0       | 0.095   | 0       |
| u18QR | 3           | <b>Lead and IQ</b>        | 0       | 0.302   | 0       |
| u21QR | 1           | <b>Concentration</b>      | 0       | 0.104   | 0       |
| u23QR | 3           | <b>Peanut Butter</b>      | 0       | 0.281   | 0       |
| u25QR | 4           | <b>Well Depth</b>         | 0       | 0.425   | 0       |
| u28QR | 2           | <b>Faulty Products</b>    | 0       | 0.288   | 0       |
| u29QR | 3           | <b>Salt Intake</b>        | 0       | 0.424   | 0       |
| u32QR | 1           | <b>Graph</b>              | 0       | 0.139   | 0       |
| u34QR | 2           | <b>Leap Years</b>         | 0       | 0.252   | 0       |
| u3CR  | 1           | <b>Bob Smith</b>          | 0       | 0       | 0.020   |
| u5CR  | 3           | <b>CR2 A</b>              | 0       | 0       | 0.319   |
| u7CR  | 1           | <b>Bennet</b>             | 0       | 0       | 0.046   |
| u8CR  | 4           | <b>Spatial/Verbal</b>     | 0       | 0       | 0.435   |
| u11CR | 3           | <b>Internet Debates</b>   | 0       | 0       | 0.263   |
| u13CR | 2           | <b>Book Club 1</b>        | 0       | 0       | 0.234   |
| u20CR | 2           | <b>Blips</b>              | 0       | 0       | 0.343   |
| u22CR | 2           | <b>CR2 B</b>              | 0       | 0       | 0.254   |
| u26CR | 4           | <b>Circuses</b>           | 0       | 0       | 0.295   |
| u27CR | 3           | <b>Camlann</b>            | 0       | 0       | 0.394   |
| u31CR | 2           | <b>Citations</b>          | 0       | 0       | 0.284   |
| u33CR | 3           | <b>Prisons</b>            | 0       | 0       | 0.278   |

The unit factor loadings are significant (based on the  $t$  statistics obtained by dividing the factor loadings with their standard errors) and almost all greater than 0.1, with most above 0.3.

Thus, most units contribute significantly to the component dimensions to which they were *a priori* assigned.

The small factor loadings are from units with only one item.

The *goodness of fit statistics*, given below, confirm that this three-factor model is a good fit to the data.

*Goodness of fit statistics:*

It is suggested in literature that in assessing goodness of fit in confirmatory factor analysis, a range of statistics should be reported as given below for this test.

1. *Goodness of fit*, GFI, is 0.945. This means that 94.5 % of the covariance in the observed data is explained by the model. "By convention, GFI should be equal to or greater than .90 to accept the model", which it is. (<http://www2.chass.ncsu.edu/garson/pa765/structur.htm>)
2. *Adjusted goodness-of-fit index*, AGFI, is 0.937. A variant of GFI which adjusts GFI for degrees of freedom, "AGFI should also be at least .90." which it is. (<http://www2.chass.ncsu.edu/garson/pa765/structur.htm>).
3. *Root Mean Square Error of Approximation* (RMSEA) is 0.0357. According to [Steiger \(1990\)](#), [Browne & Cudeck, \(1993\)](#), RMSEA deals with the fit of the model to the population covariance matrix. Values of RMSEA less than 0.05 indicate good fit, which is the case. ([MacCallum, Browne, & Sugawara, 1996](#)).
4. *The Standardised Root Mean Square Residuals* (SRMR) is 0.0371. "This should be .06 or less to indicate a good fit.", which it is. (<http://www.bangor.ac.uk/%7Epes004/resmeth/lisrel/cfa.htm>)
5. *Hoelter's Critical N* is the size the sample must be for the researcher to accept the model by chi-square, at the 0.05 or 0.01 levels. Critical N here is about 700. It is suggested that Critical N should be greater than 200, which it is. (<http://www2.chass.ncsu.edu/garson/pa765/structur.htm>).
6. *Comparative fit index*, CFI, compares the existing model fit with a null model which assumes the latent variables in the model are uncorrelated (the "independence model"). CFI for this test is 0.920. "By convention, CFI should be equal to or greater than .90 to accept the model.", which it is. (<http://www2.chass.ncsu.edu/garson/pa765/structur.htm>).
7. *Incremental fit index*, IFI, is also known as *DELTA*. IFI for this test is 0.920. IFI should be equal to or greater than 0.90 to accept the model, which it is. (<http://www2.chass.ncsu.edu/garson/pa765/structur.htm>).

All of this suggests that the three factor solution fits the data well.

Confirmatory factor analysis at the five factor level using the three CR factors separately, as well as QR and VP, slightly improved the goodness of fit. Differential use of the sub-components of CR might therefore be considered in future.

This all suggests that the test may be usable in either two or three factor format, or that the CR sub-components might be used differentially, but empirical work needs to be done to identify the optimal way to use the test items to predict success in various university courses (for example, the optimal weighting of component scores to maximise prediction for particular courses).

## Measurement Error

In the current test, the standard error of measurement is smallest in the logit interval from -2 to 2. Over this interval, the standard error for each 30-item scale is around 0.5 logit. Above 2 logits, the measurement error is greater.

However, given that it is likely that for any student at least two of the test components (60 items) will be used, or the three components will all be used (perhaps differentially weighted), the actual measurement error for a student will be significantly less than this.

Even though test reliability is generally satisfactory, consideration needs to be given as to whether the level of measurement error is appropriate for the actual test usage. Since measurement error is least when item difficulties and student abilities match, it may be that more items are required in certain regions of the scale for certain uses, especially around critical cut-off/selection levels, or two or more versions of the test are needed for candidate groups with different ability ranges.